ISPRS Journal of Photogrammetry and Remote Sensing 145 (2018) 44–59

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

Building instance classification using street view images

Jian Kang [a], Marco Körner [b], Yuanyuan Wang [a], Hannes Taubenböck [c], Xiao Xiang Zhu [a,d,*]

[a] Signal Processing in Earth Observation (SiPEO), Technical University of Munich (TUM), 80333 Munich, Germany
[b] Chair of Remote Sensing Technology, Technical University of Munich (TUM), 80333 Munich, Germany
[c] German Remote Sensing Data Center (DFD) (IMF), German Aerospace Center (DLR), 82234 Wessling, Germany
[d] Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), 82234 Wessling, Germany

ARTICLE INFO

Article history:
Received 9 June 2017
Received in revised form 8 November 2017
Accepted 7 February 2018
Available online 2 March 2018

Keywords:
CNN
Building instance classification
Street view images
OpenStreetMap

ABSTRACT

Land-use classification based on spaceborne or aerial remote sensing images has been extensively studied over the past decades. Such classification is usually a patch-wise or pixel-wise labeling over the whole image. But for many applications, such as urban population density mapping or urban utility planning, a classification map based on individual buildings is much more informative. However, such semantic classification still poses some fundamental challenges, for example, how to retrieve fine boundaries of individual buildings. In this paper, we proposed a general framework for classifying the functionality of individual buildings. The proposed method is based on Convolutional Neural Networks (CNNs) which classify façade structures from street view images, such as Google StreetView, in addition to remote sensing images which usually only show roof structures. Geographic information was utilized to mask out individual buildings, and to associate the corresponding street view images. We created a benchmark dataset which was used for training and evaluating CNNs. In addition, the method was applied to generate building classification maps on both region and city scales of several cities in Canada and the US.

1. Introduction

The classification of land cover from Earth Observation (EO) images in complex urban environments has been a focus in remote sensing over the past decades (Anderson et al., 1976; Pal and Mather, 2003; Yuan et al., 2005; Stefanov et al., 2001; Rodriguez-Galiano et al., 2012; Albert et al., 2017). Beyond, high resolution spaceborne and aerial images are one of the handful information sources for monitoring urban development on large scales.

However, the transfer from land cover to land use in EO-data is complex and relies mostly on the geometry and the appearance of individual buildings and the patterns they group together (Lu and Weng, 2006; Gong et al., 1992; Paola and Schowengerdt, 1995; Pacifici et al., 2009; Khorram et al., 1987; Di et al., 2000; Cheng et al., 2015; Huang et al., 2014, 2015, 2017). The correlation of physical indicators such as building volumes, density or alignment has been used to infer the usage of buildings, e.g. as commercial areas (e.g. Fig. 1(a)), residential areas (e.g. Fig. 1(b)) or industrial areas (e.g. Fig. 1(c)). Nevertheless, such pattern analysis cannot be directly transferable to the classification of individual buildings as we go to a finer level of urban intrinsic scale. For example, Fig. 1 (a) shows a commercial area comprised of multiple high-rise buildings. However, the label "commercial area" cannot be assigned to all the building instances within it. As illustrated in Fig. 2, the corresponding street view images show that the commercial area is comprised of a few apartments, office buildings, and one church. This also applies to the example shown in Fig. 1(b) and (c), where both the residential and industrial areas are comprised of buildings with different functionalities. As can be seen, land-use classification at a level of individual buildings is not a trivial task. Usually, such a classification map is only obtainable through city cadastral databases, not accessible or sometimes even not existent. Updating such databases without automatic methods can be very labor intensive. Hence, automatically achieving a building instance-level classification is necessary and can be beneficial for applications related with urban planning. Towards an automatic classification of individual buildings, the challenges are twofold. Firstly, remote sensing images usually only contain roof structures due to their nadir-looking imaging geometry. The visual difference of the roofs between certain building classes, e.g. apartments and office buildings, can be subtle, as an example shown in Fig. 2. Secondly, the extraction of building footprints directly from

* Corresponding author at: Signal Processing in Earth Observation, Technical University of Munich, Arcisstr. 21, 80333 Munich, Germany and Remote Sensing Technology Institute, German Aerospace Center, 82234 Wessling, Germany.
E-mail address: xiao.zhu@dlr.de (X.X. Zhu).

https://doi.org/10.1016/j.isprsjprs.2018.02.006

Fig. 1. Examples of land-use classification.

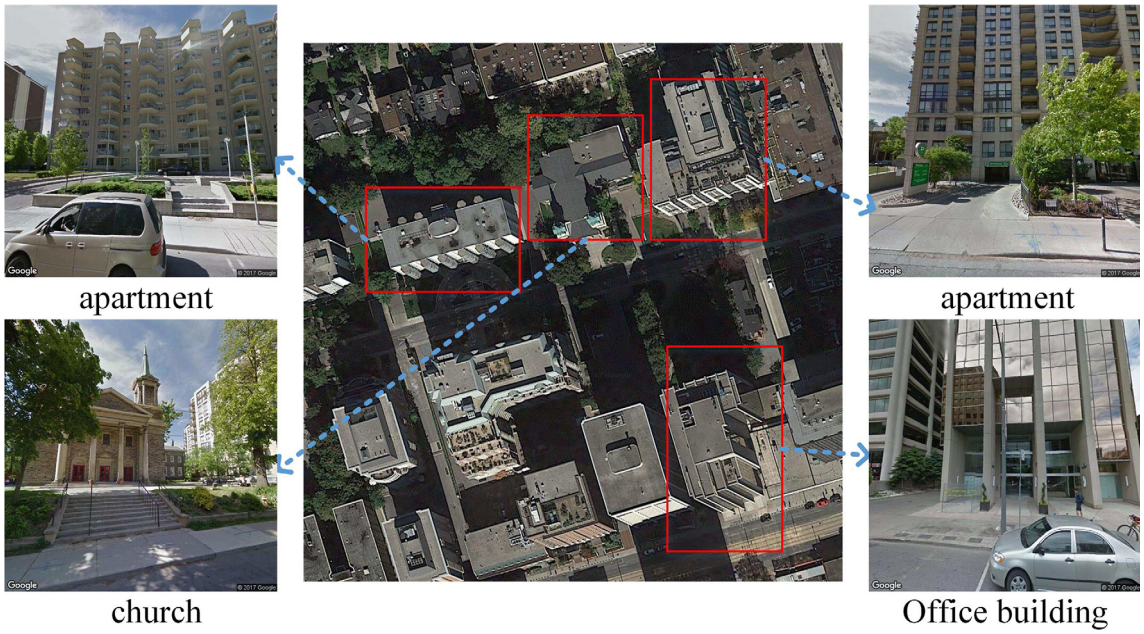

Fig. 2. The commercial land-use area as shown in Fig. 1(a), along with the street view images for some buildings selected by the red rectangles. These buildings do not belong to the same category, even though they are located in the same land-use area. Besides, compared to the roof structures, the information of façade structures displayed in street view images is richer and more sufficient to be used for building instance classification.

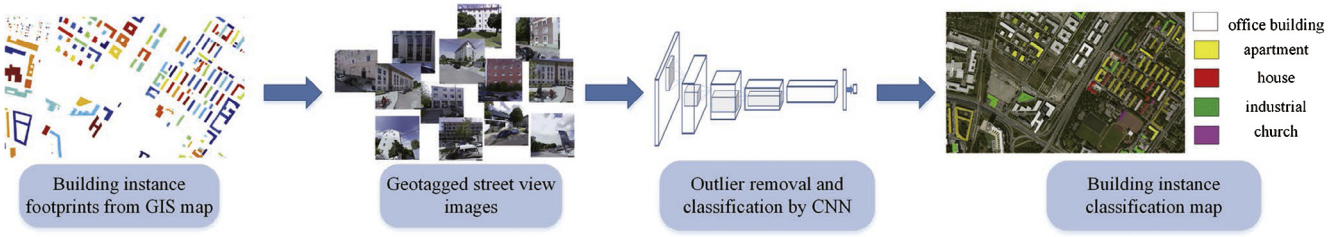

Fig. 3. The proposed workflow for land-use classification at a level of individual buildings.

remote sensing images is still under preliminary research. A clear segmentation of building footprints usually requires height information which comes at an additional cost.

In this paper, we propose a general framework to tackle the abovementioned challenges, which exploits the information extraction from freely available street view images and online geographic maps. Specifically, façade structures shown in online street view images are sufficiently rich for building functionality classification, and the online map services, such as OpenStreetMap (OpenStreetMap, 2017) or Google Maps, can provide the building footprints which can be associated to street view images via their geographic locations. As shown in Fig. 2, the façades displayed in

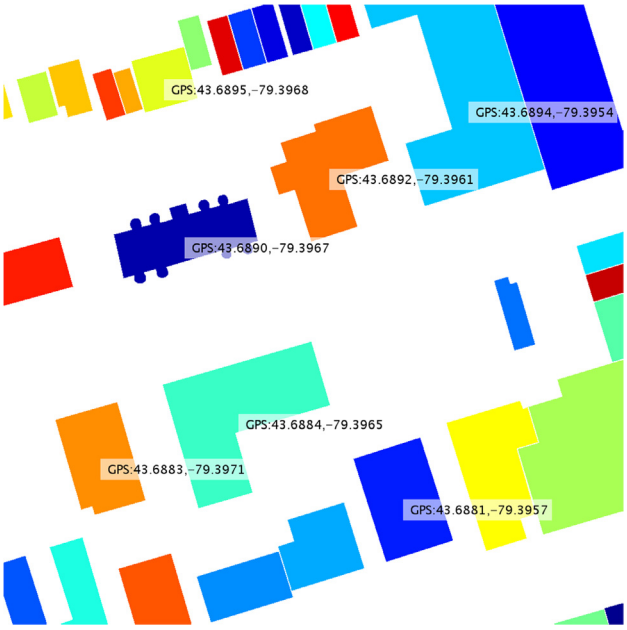

Fig. 4. Geographic information (GPS) retrieved from Google Maps of the remote sensing image in Fig. 2, with the color randomly assigned to each building mask.

street view images reveal much more details of different types of buildings than the corresponding roof patches. Therefore, building instances are classified based on their geo-tagged street view images in the proposed method, and the inferred labels are then linked to individual building footprints through spatial clustering. We also build a benchmark dataset of building street view images to train Convolutional Neural Networks (CNNs) for the classification over large areas, as CNN has been demonstrated its powerful ability in the tasks of this sort (Russakovsky et al., 2015; Zhou et al., 2016; Lin et al., 2014).

In a summary, the contributions of this paper are listed as follows:

- Proposed a general framework for land-use classification at a level of individual buildings.
- Built a street view benchmark dataset for training building instance CNN classifiers based on façade structures. The dataset utilized in this paper can be downloaded via www.sipeo.bgu.tum.de/downloads/BIC_GSV.tar.gz
- The obtained building classification maps demonstrated the potentials for many innovative urban analysis, e.g. very high resolution urban population density mapping, urban social structure understanding, city economy structure analysis and general urban planning.

2. Related work

Feature extraction from remote sensing images plays a vital role in land-use classification. Handcrafting features have been well studied for decades, such as scale-invariant feature transform (SIFT) (Lowe, 1999) encoded by bag of visual words (BoVW) (Yang and Newsam, 2010; Cheriyadat, 2014; Zhu et al., 2016), multiple textural features (Xu et al., 2016), 3D features derived from a digital surface model (Taubenböck et al., 2013) and features learned by sparse coding methods (Wang et al., 2014; Yang et al., 2014; Sun et al., 2015; Rigas et al., 2013; Zhang et al., 2015; Tuia et al., 2015, 2016; Yao et al., 2016; Cheng et al., 2015).

Recently, many approaches based on deep learning techniques have emerged (Cheng et al., 2017; Ma et al., 2016; Zhang et al., 2018). Chen et al. (2014) proposes a hierarchical feature extraction method via stacked autoencoders, which merges both spectral and spatial information of hyperspectral images for land-use classification. In Zou et al. (2015), deep belief networks are employed for the feature learning in remote sensing scene classification. Both Penatti et al. (2015) and Marmanis et al. (2016) investigate the possibility of transferring features learned by CNN from *ImageNet* dataset (Deng et al., 2009) to achieve remote sensing image classification by fine-tuning procedures. To improve the composition-based inference of land-use classes, multiscale CNN-based approaches are developed in Zhao and Du (2016), Luus et al. (2015), and Liu et al. (2016). By exploiting deep Boltzmann machine, a novel weakly supervised learning approach for object detection in remote sensing images is introduced (Han et al., 2015). For effectively dealing with the problem of object rotation variations, a rotation-invariant CNN model is proposed in Cheng et al. (2016). Based on greedy layerwise unsupervised pretraining, Romero et al. (2016) proposes a novel unsupervised deep feature extraction method. Taking advantage of geographical information from OpenStreetMap, a fully convolutional neural network is trained to achieve pixel-wise classifications in optical images on large scales (Maggiori et al., 2017). Recurrent Neural Network (RNN) is also proved to be efficient for classifying sequence-based data like hyperspectral images (Mou et al., 2017). An end-to-end fully Conv-Deconv network for unsupervised spectral-spatial feature extraction in hyperspectral images has been proposed in Mou

Fig. 5. Outlier examples of the retrieved street view images. We can see that there is no available information of building façades for the classification.

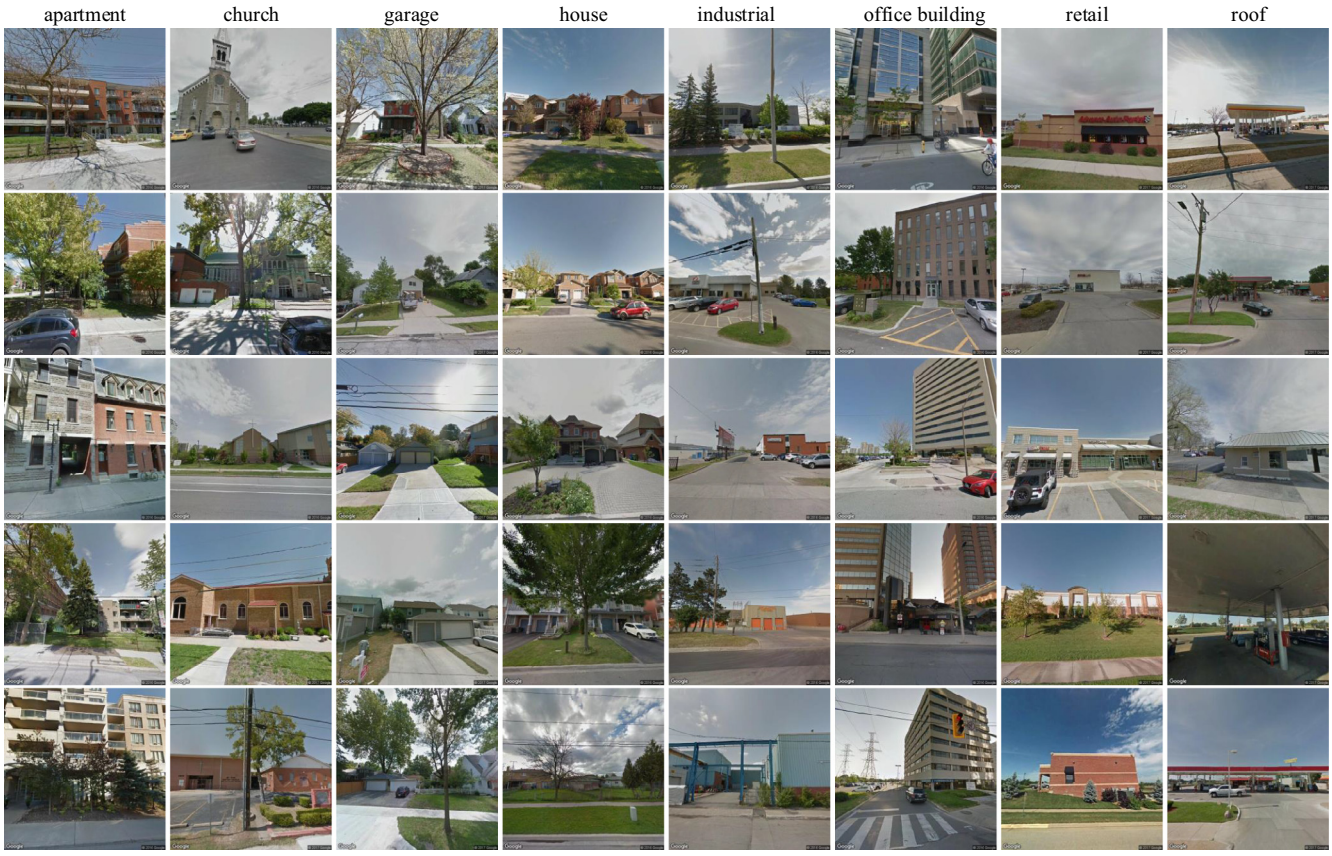

Fig. 6. Examples of the benchmark dataset. It totally contains 19,658 street view images of buildings with eight classes, i.e. *apartment, church, garage, house, industrial, office building, retail and roof*. The images are downloaded from Google StreetView (Anguelov et al., 2010), and the associated labels are jointly retrieved from OpenStreetMap based on the geographic information.

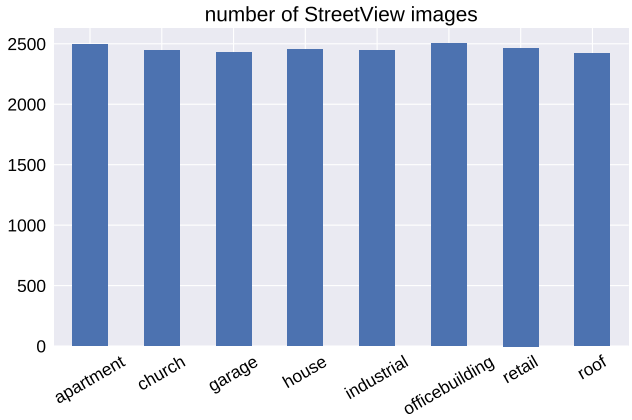

Fig. 7. Number of street view images of each building class.

et al. (2018). In order to better interpret land-uses of Synthetic Aperture Radar (SAR) images in urban areas, Hughes et al. (2018) proposes a pseudo-siamese CNN for identifying corresponding patches in very-high-resolution (VHR) optical and SAR remote sensing imagery. Surveys about the applications of deep learning techniques to land-use classification with remote sensing images are proposed in Zhang et al. (2016), Zhu et al. (2017).

Even the abovementioned literature is of course not exhaustive, none of them have explicitly addressed the land-use classification at a level of individual buildings.

3. Overall workflow

As illustrated in Fig. 3, the proposed workflow for building instance classification contains the following steps:

- Retrieval of building footprints and associated street view images.
- Outlier removal by the pretrained CNN on Places2 dataset (Zhou et al., 2016).
- Building instance classification by the CNN trained on our benchmark dataset.

3.1. Retrieval of building footprints and street view images

The building footprints and their geographic locations, can be retrieved from online geographic information systems (GIS), such as OpenStreetMap or Google Maps. For example, the building footprints of the area shown in Fig. 2 are displayed in Fig. 4, along with the associated GPS coordinates (latitude, longitude). The color is randomly assigned to indicate different building instances. Given these GPS coordinates, we can download the corresponding Google StreetView images (Anguelov et al., 2010) which show façade structures of individual buildings, since the retrieved images can display these specific locations by the closest panoramas.

3.2. Outlier removal by pretrained CNN on Places2 dataset

Due to the uncontrolled quality of street view images, many of them cannot be directly utilized for the building classification. For

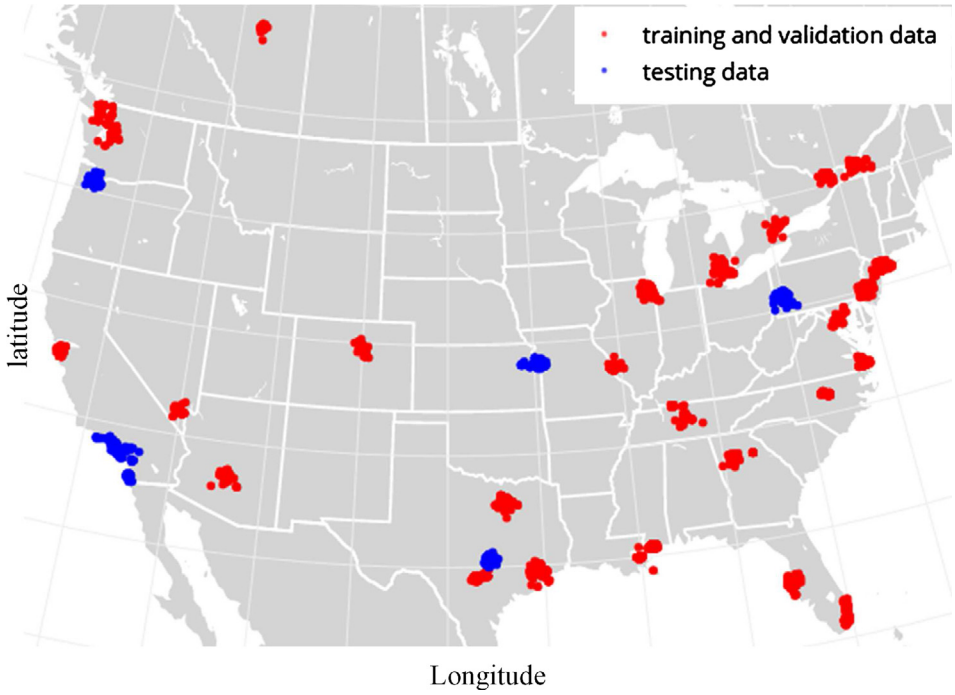

Fig. 8. GPS locations of our benchmark dataset. We split all the images into two parts: one for training (17,600 images) and the others for testing (2058 images). Note that all the testing images are located in different cities with the training ones.

Table 1
Building class descriptions from OpenStreetMap.

Apartment	A building arranged into individual dwellings, often on separate floors. May also have retail outlets on the ground floor
Church	A building that was built as a church
Garage	A building suitable for the storage of one or possibly more motor vehicle or similar
House	A dwelling unit inhabited by a single household (a family or small group sharing facilities such as a kitchen)
Industrial	A building where some industrial process takes place
Office building	A building where non-specific commercial activities take place
Retail	A building primarily used for selling goods that are sold to the public
Roof	A structure that consists of a roof with open sides, such as a rain shelter, and also gas stations

example, as shown in Fig. 5, one retrieved image is taken from the building interior and the other two buildings are occluded by a vehicle and trees on the side-walks. Therefore, the corresponding façade structures are not available for classifying these buildings. These outliers can severely influence the classification results. For removing them, we employ the released VGG16 model (Simonyan and Zisserman, 2014) trained on Places2 dataset (Zhou et al., 2016) to preliminarily screen the street view images, as this architecture has achieved the highest top-1 accuracy.[1] The dataset contains almost 10 million scene photos, labeled with 476 scene categories and attributes, which include the building-related categories, i.e. [*apartment, church, house, industrial area, museum, building facade, embassy, hospital, parking garage, hotel*]. Only the images belonging to the abovementioned categories are preserved for the follow-up classification.

3.3. Building instance classification

To train a building instance classifier, we first build a corresponding street view benchmark dataset, which contains totally 19,658 images from eight classes, i.e. *apartment, church, garage, house, industrial, office building, retail and roof*, and there are around 2500 images for each building class, as shown in Figs. 6 and 7. The geo-tagged images are downloaded through Google StreetView API,[2] with the associated metadata,[3] i.e. the image size and pitch value are set to be 512×512 pixels and 10degrees, respectively. As illustrated in Fig. 8, all the street view images are located over several cities of the US and Canada, e.g. Montreal, New York and Denver, and their associated ground truth building labels are extracted from OpenStreetMap.[4] The descriptions for the building classes are demonstrated in Table 1.

Since the dataset is not sufficiently large to train a CNN with millions of parameters from the scratch, we choose to fine-tune a pretrained CNN with our dataset. It is common that a pretrained CNN on a large dataset such as *ImageNet* (Russakovsky et al., 2015) can be well adapted to other new tasks with small scale datasets, since low-level features such as corners and edges generated by prior layers of CNN are general in different images. The high-level image representations extracted by posterior layers are dependent on different tasks. Therefore, fine-tuning the layers of the pretrained CNN with the new dataset has been proven to be an efficient way for the adaptation of the CNN to a new training task.

To further improve the classification robustness, the street view images for each building instance are classified, and the building class can be obtained in a decision level. Assuming there are Mstreet view images retrieved of the study building instance, the final building class $y$can be determined by

[1] https://github.com/metalbubble/places365.

[2] https://developers.google.com/maps/documentation/streetview/.

[3] https://developers.google.com/maps/documentation/streetview/intro.

[4] http://wiki.openstreetmap.org/wiki/Map_Features#Building.

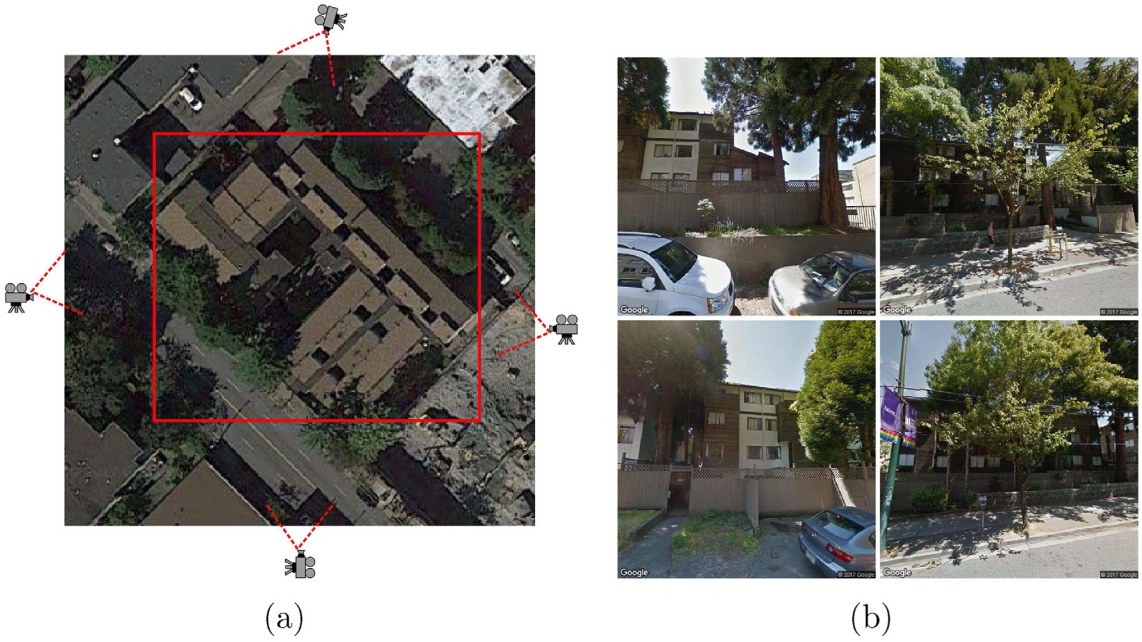

Fig. 9. (a) Illustration of different looking-angles for the same building (red rectangular). (b) The corresponding retrieved street view images: Left column shows the obvious building façades, while the right two images are outliers. In order to improve the robustness, we classify several street view images for one building, and fuse their classification labels in a decision level. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

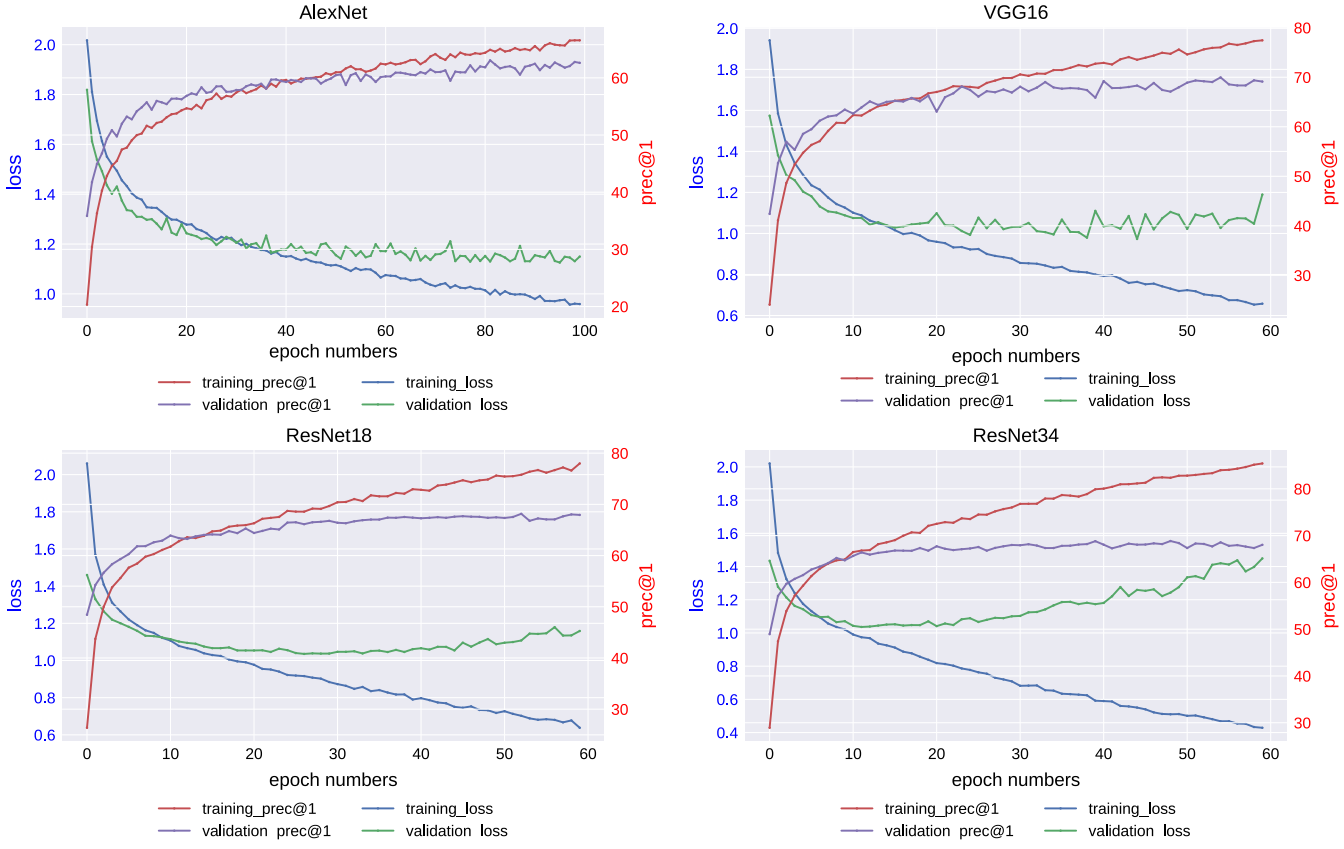

Fig. 10. The learning and top 1-precision curves of the four networks, i.e. AlexNet (Top-left), VGG16 (Top-right), ResNet18 (Bottom-left) and ResNet34 (Bottom-right). It can be seen that training losses of the four networks reduce as the epochs increase. Besides, the validation learning curve of AlexNet converges until 80 epochs, and those of the other three networks can converge within 60 epochs. Overfitting behaviors are found in ResNet18 and ResNet34, and it is more severe in ResNet34. One plausible reason is that the total parameter number of ResNet34 (21 million) is more than that of ResNet18 (11 million). As shown by top-1 precisions, AlexNet can achieve about 65%, while the other networks can obtain about 70%.

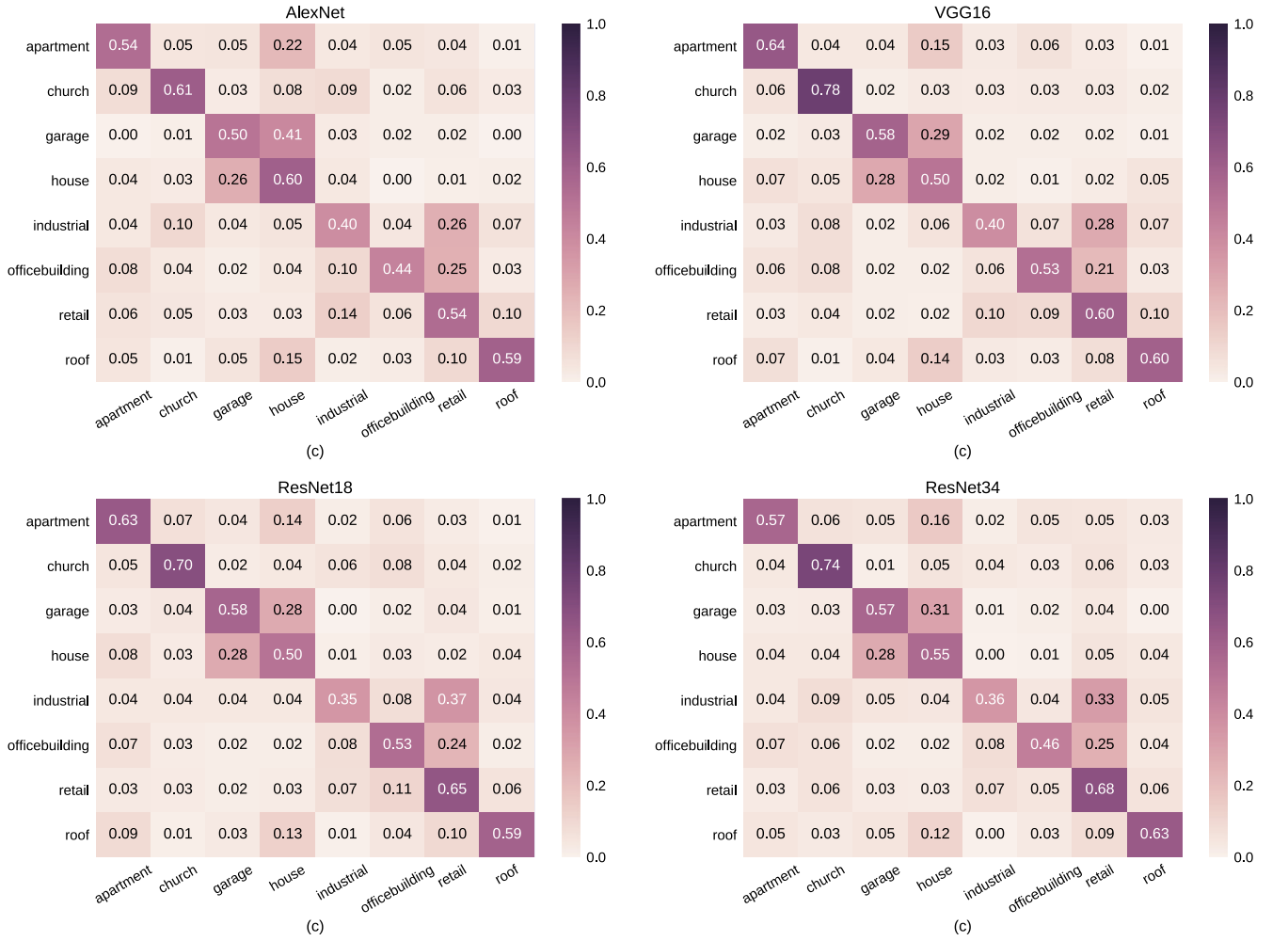

Fig. 11. The associated normalized confusion matrices of the four networks evaluated on the test images, i.e. AlexNet (Top-left), VGG16 (Top-right), ResNet18 (Bottom-left) and ResNet34 (Bottom-right).

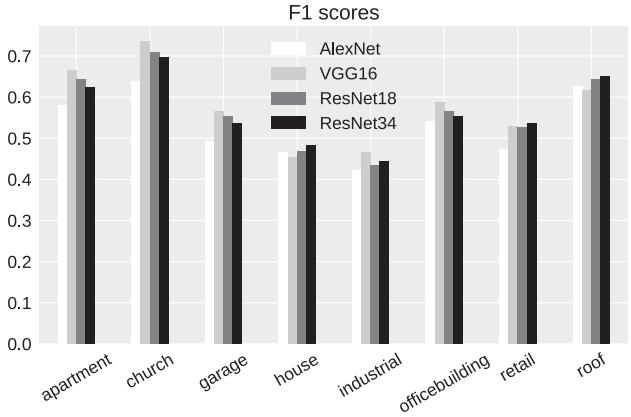

Fig. 12. F1 score performances of the four trained networks on the eight building classes. For the classes of apartment, church, garage, industrial and office building, VGG16 achieves the highest F1 score, and for the other classes, ResNet34 is the best among them.

$$y = \underset{i}{\operatorname{argmax}} \frac{1}{M} \sum_{j=0}^{M-1} f_i^{(j)}, \tag{1}$$

where $f_i^{(j)}$ is the ith element of the CNN *softmax* layer output $\mathbf{f}^{(j)}$, which denotes the probability distribution over the whole building classes, and $j$is the index of the classified street view image. For example, Fig. 9 shows a building to be classified and its corresponding street view images from four different positions. After the right two images filtered by the outlier removal step, we can calculate the final probability distribution vector by averaging those of the left two images and obtain the building class accordingly.

Table 2
Overall precisions, recalls and F1 scores.

Network	Precision	Recall	F1 score
AlexNet	0.55	0.53	0.53
VGG16	**0.59**	**0.58**	**0.58**
ResNet18	0.58	0.57	0.57
ResNet34	**0.59**	0.57	0.56

Bold values refer to the highest performance that the listed CNNs achieved.

4. Experiments

We train several state-of-the-art CNN architectures, e.g. AlexNet (Krizhevsky et al., 2012), VGG (Simonyan and Zisserman, 2014) and ResNet (He et al., 2016) by fine-tuning all the convolutional layers with our benchmark dataset, and demonstrate the corresponding training and testing performances. Among those networks, we choose the best one for generating building classification maps both on region and city scales.

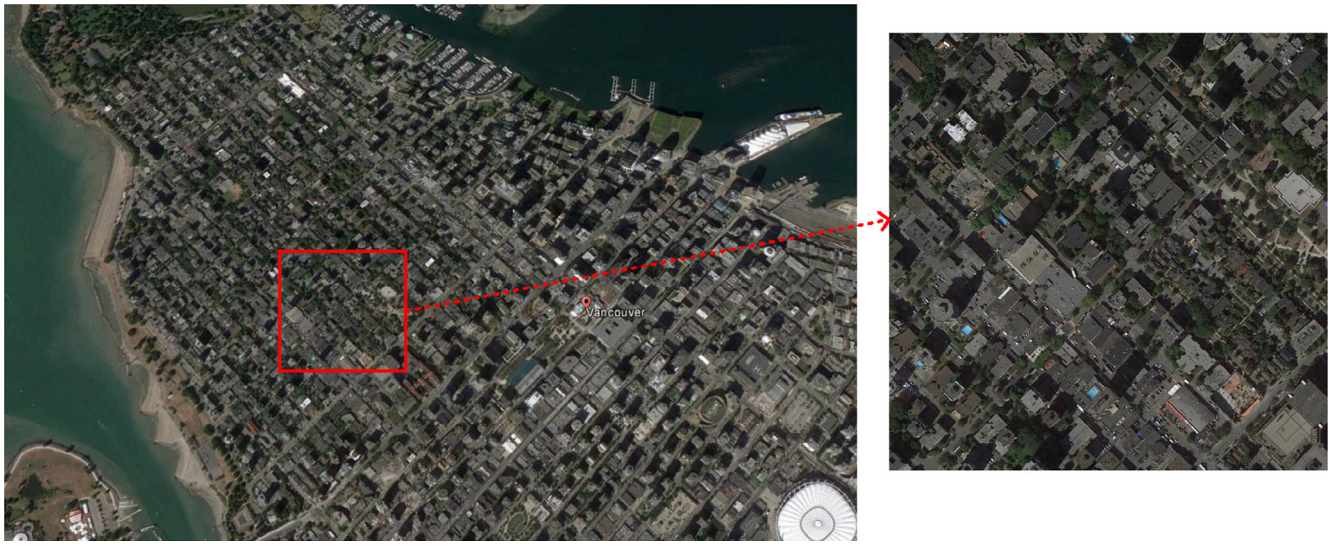

Fig. 13. Illustration of one study area in Vancouver (image is from Google Earth).

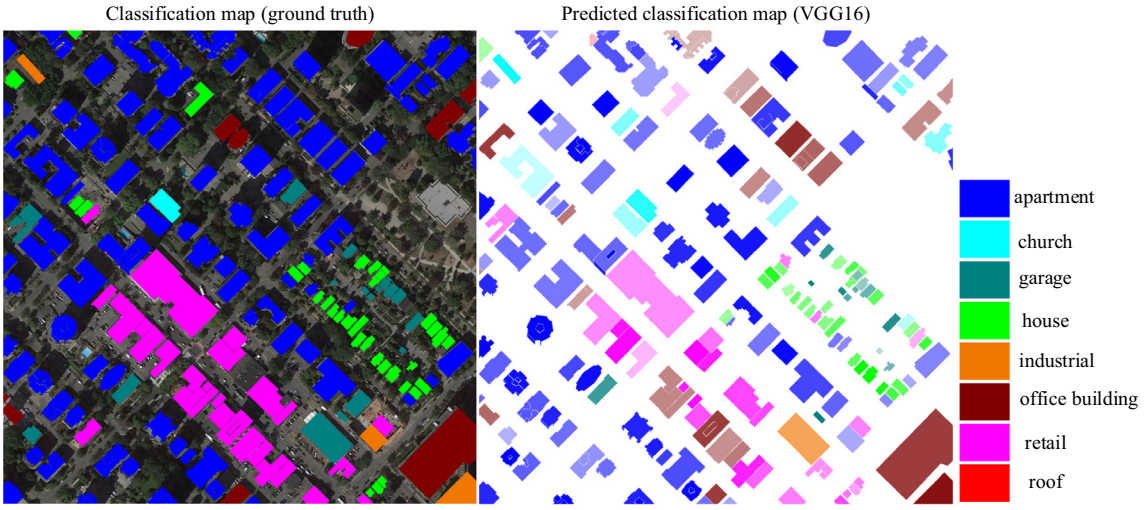

Fig. 14. The predicted building classification map (Right), along with the ground truth (Left), where different colors represent different building classes. The total number of building instances in this area is 196. Our result predicts 93 apartments, 10 churches, 13 garages, 24 houses, 1 industrial building, 21 office buildings, 26 retails and 1 roof. 7 buildings are not classified, since no corresponding street view images are found. Moreover, the confidence score for the class of each building is shown by the opacity of the associated color mask, i.e. the higher the opacity, the larger the confidence score and vice versa.

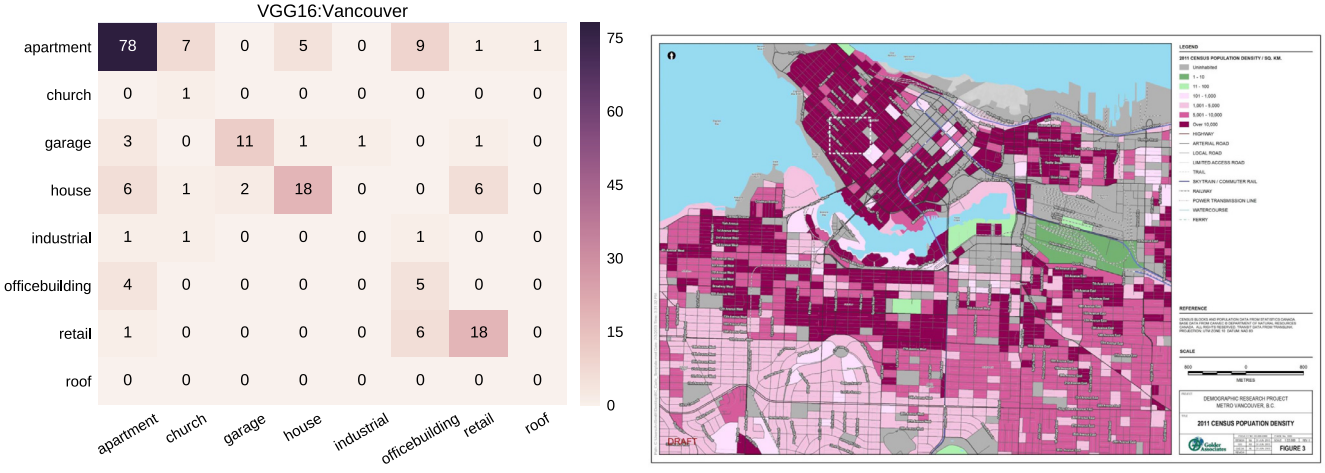

Fig. 15. (Left) The confusion matrix of the classification result of the area in Vancouver. We can see that this area is mainly composed by apartments. (Right) 2011 census population density of Vancouver. The white rectangle indicates the study area, which has a high population density of over 10,000/km^2.

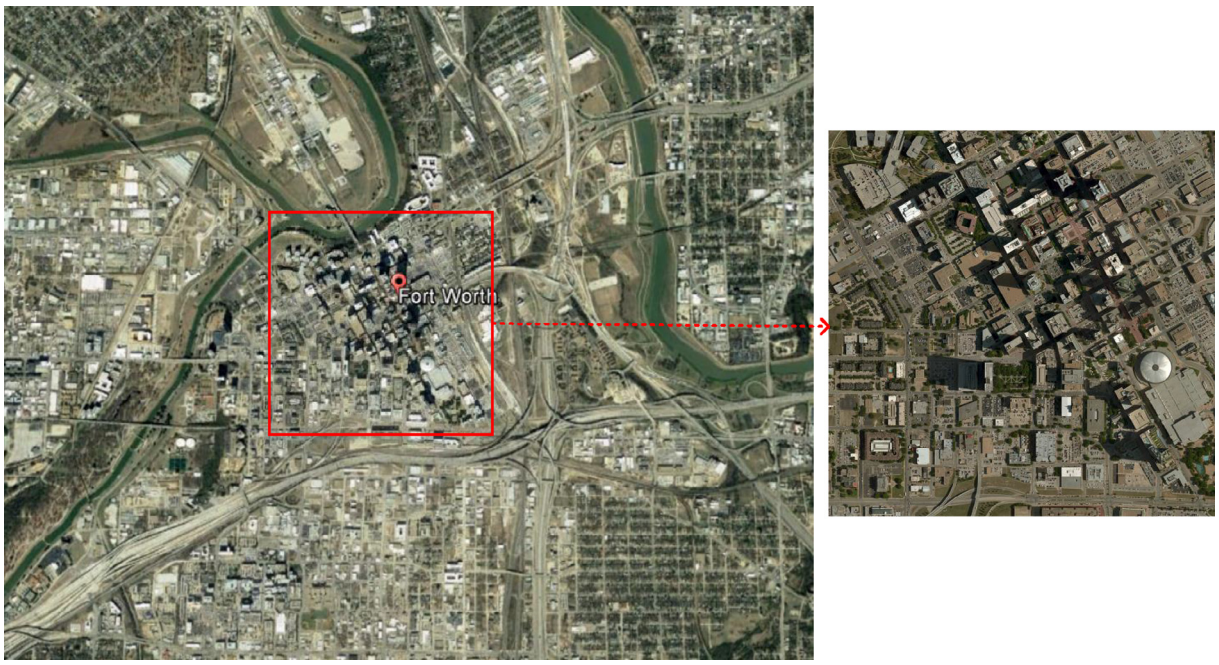

Fig. 16. Illustration of one study area in Fort Worth (image is from Google Earth).

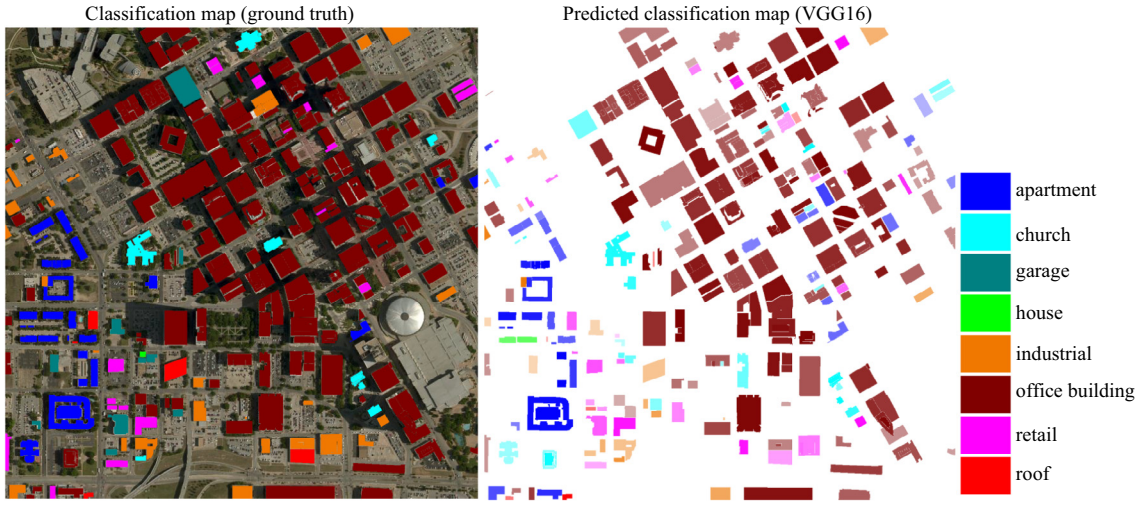

Fig. 17. The predicted building classification map (Right), along with the ground truth (Left). The total number of buildings in this area is 316. Our result predicts 34 apartments, 30 churches, 2 houses, 19 industrial buildings, 152 office buildings, 28 retails and 6 roofs. There are no street view images for the remaining 45 buildings.

4.1. Training

As illustrated in Fig. 8, we split the whole dataset into two parts: 17,600 images for training (2200 images for each building class) and 2058 images for testing. Note that all the testing images are retrieved from different cities with those utilized for training. In order to monitor the training status of networks, we randomly select 3200 images from the training samples to be the validation data. We train four different networks i.e. AlexNet, VGG16, ResNet18 and ResNet34 following the same procedure: Convolutional layers of all these networks are initialized by those pre-trained with *ImageNet*, and fully connected layers are randomly initialized following a uniform distribution.

Each training batch contained in total 64images. The stochastic gradient descent algorithm with a learning rate of $\eta = 5 \cdot 10^{-4}$ and a momentum value of $p = 0.9$ was employed for training. To adjust the learning rate, we decayed its value by a factor of 0.1 in every 30epochs. Cross-entropy loss was utilized for training with the weight decay parameter of $w = 10^{-5}$. The neurons of fully connected layers were dropped out by a probability of 25%. To augment the training data, we randomly cropped 224×224 pixels from the original 256×256 pixels and randomly flipped the cropped images horizontally. All the experiments were implemented with Pytorch[5] and carried out by one NVIDIA TITAN X (Pascal) 12 GB GPU.

As shown in Fig. 10, we plot the learning curves of both training and validation data, and calculate the corresponding top 1-precision values during training. It can be seen that training losses of the four networks reduce as the number of epochs increases.

[5] http://pytorch.org/.

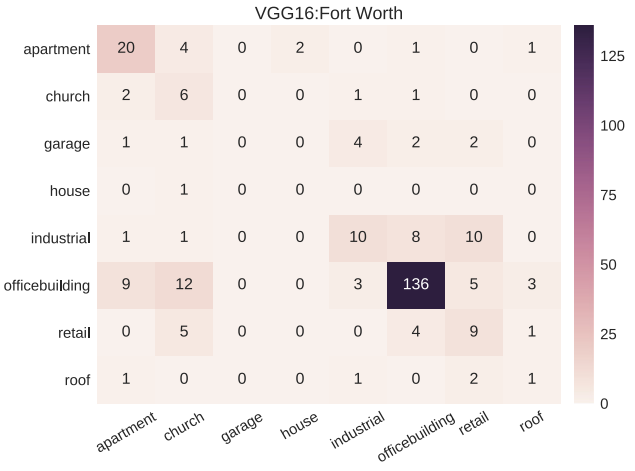

Fig. 18. The confusion matrix of the classification result of the area in Fort Worth. We can see that this area is mainly composed by office buildings, which indicates that it is a business district and may locate in the center of Fort Worth.

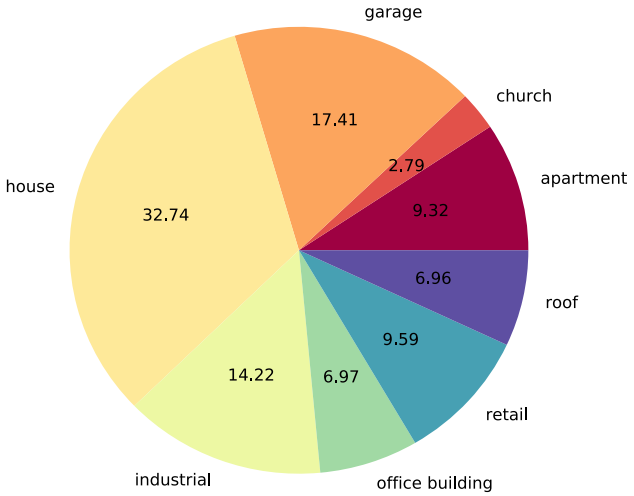

Fig. 20. Pie chart of the building class proportions of the predicted buildings of Calgary.

Besides, the validation learning curve of AlexNet converges until 80 epochs, and those of the other three networks can converge within 60 epochs. Overfitting behaviors are found in ResNet18 and ResNet34, and it is more severe in ResNet34. One plausible reason is that the total parameter number of ResNet34 (21 million) is more than that of ResNet18 (11 million). As shown by the top-1 precisions, AlexNet can achieve about 65%, while the other networks can obtain about 70%. For the follow-up evaluations, we choose ResNet18 trained until 40 epochs and 25 epochs of ResNet34 and compare the performances of those four networks.

4.2. Testing

As illustrated in Figs. 11 and 12, we demonstrate the normalized confusion matrices of all the trained networks evaluated by our test data, and the associated F1 scores of the eight building classes, respectively. F1 score (F_1), also known as F-measure, is a criteria to measure classification accuracy, which considers both the precision $p$and the recall r. It is defined as

$$F_1 = 2 \cdot \frac{p \cdot r}{p + r}. \tag{2}$$

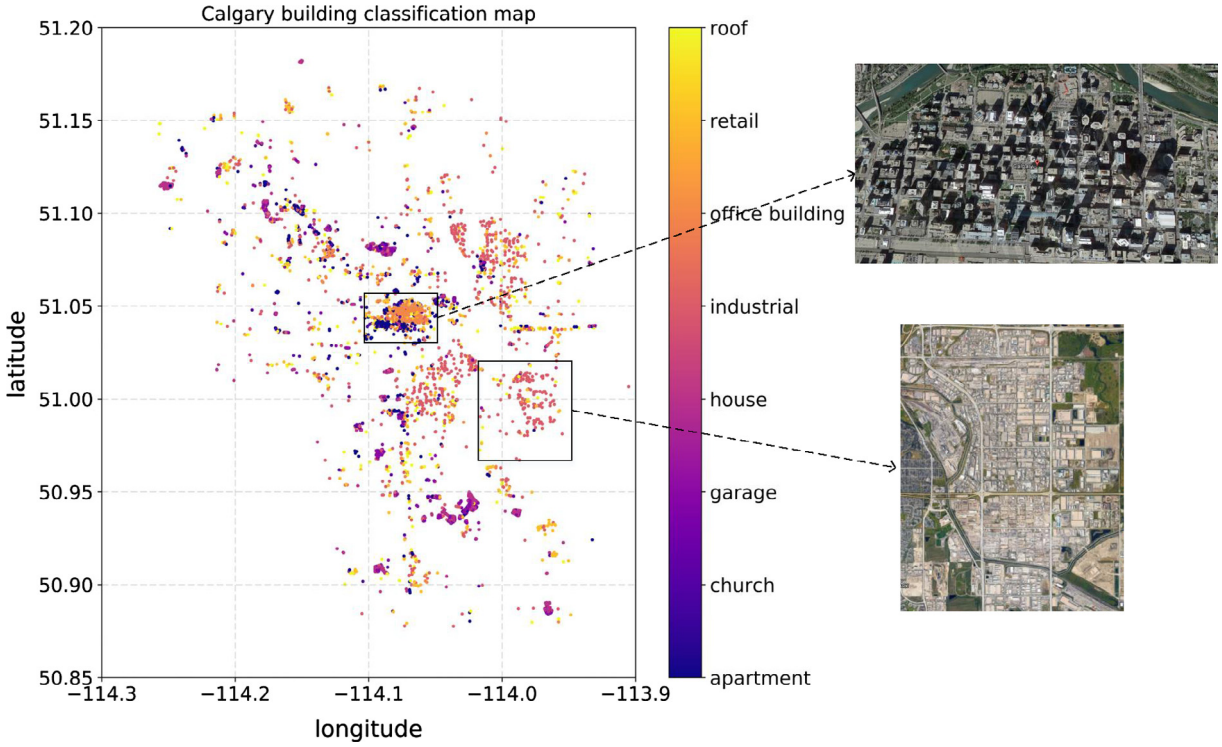

Fig. 19. The city-scale building classification map of Calgary, where each classified building instance is displayed as a colored point with GPS coordinates. It is obvious that there are three main industrial districts, and the downtown area is crowded by office buildings. Correspondingly, we also present the remote sensing images of one industrial and the downtown areas (black rectangles). Such classification map can infer that Calgary is an industry city with single central business district to which the three main industrial blocks are located. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 3
Classification performance of randomly selected 1000 buildings of Calgary.

	Precision	Recall	F1 score	Support
Apartment	0.54	0.77	0.64	56
Church	0.00	0.00	0.00	1
Garage	0.41	0.90	0.57	124
House	0.97	0.62	0.75	630
Industrial	0.51	0.80	0.63	82
Office building	0.65	0.19	0.29	58
Retail	0.33	0.37	0.35	43
Roof	0.15	0.83	0.25	6
Overall	0.78	0.64	0.66	1000

Table 4
Classification performance of randomly selected 1000 buildings of Boston.

	Precision	Recall	F1 score	Support
Apartment	0.35	0.42	0.38	137
Church	0.06	0.80	0.11	5
Garage	0.51	0.38	0.43	221
House	0.69	0.61	0.65	546
Industrial	0.07	0.25	0.11	4
Office building	0.58	0.62	0.60	60
Retail	0.20	0.42	0.27	19
Roof	0.62	0.62	0.62	8
Overall	0.58	0.53	0.55	1000

Table 5
Classification performance of randomly selected 1000 buildings of Toronto.

	Precision	Recall	F1 score	Support
Apartment	0.73	0.83	0.78	212
Church	0.29	0.59	0.39	22
Garage	0.18	0.42	0.25	33
House	0.94	0.73	0.82	575
Industrial	0.36	0.79	0.49	24
Office building	0.04	0.25	0.06	4
Retail	0.84	0.50	0.63	117
Roof	0.33	0.92	0.49	13
Overall	0.82	0.71	0.75	1000

Moreover, the overall precisions, recalls and F1 scores of the four networks are demonstrated in Table 2. From the results, we can see that the classification performance of AlexNet is worse than the other three networks. For the classes of apartment, church, garage, industrial and office building, VGG16 achieves the highest F1 score, and for the other classes, ResNet34 is the best among them. According to the overall accuracies shown in Table 2, we choose the trained VGG16 model for the upcoming generation of building classification maps of the study areas.

4.3. Building classification maps of study areas

4.3.1. Maps of study areas in Vancouver and Fort Worth

One testing area in Vancouver (image is from Google Earth) can be seen in Fig. 13. The associated ground truth and predicted building classification maps are present in Fig. 14, where different colors represent different building classes. We also draw the corresponding confusion matrix of the inferred result in Fig. 15. The total number of building instances in this area is 196. Our result predicts 93 apartments, 10 churches, 13 garages, 24 houses, 1 industrial building, 21 office buildings, 26 retails and 1 roof. 7 buildings are not classified, since no corresponding street view images are found. Moreover, the confidence score for the class of each building is shown by the opacity of the associated color mask, i.e. the higher the opacity, the larger the confidence score and vice versa. From the results, we can see that this study area is mainly composed of apartments, which indicates that it is a residential district and there may be a high population density of this area. Our analysis is confirmed by the 2011 census population density of Vancouver downloaded from the website,[6] as shown in Fig. 15(Right). The white rectangle in the figure marks the study area which has the highest population density (over $10,000/km^2$). Such classification map gives an insight into the social structure of an residential area. For example, the houses and retails are both grouped together at the right corner of this district.

Another testing area is located in Fort Worth, shown by the red rectangle area in Fig. 16. The ground truth and predicted building classification maps are present in Fig. 17. The associated confusion matrix is demonstrated in Fig. 18. The total number of buildings in this area is 316. Our result predicts 34 apartments, 30 churches, 2 houses, 19 industrial buildings, 152 office buildings, 28 retails and 6 roofs. There are no street view images for the remaining 45 buildings. According to the predicted result, we can see that this area is mainly composed of office buildings, which indicates that it is a business district and may locate in the center of Fort Worth.

[6] https://blogs.ubc.ca/maps/2013/07/03/vancouverpopulationdensity/.

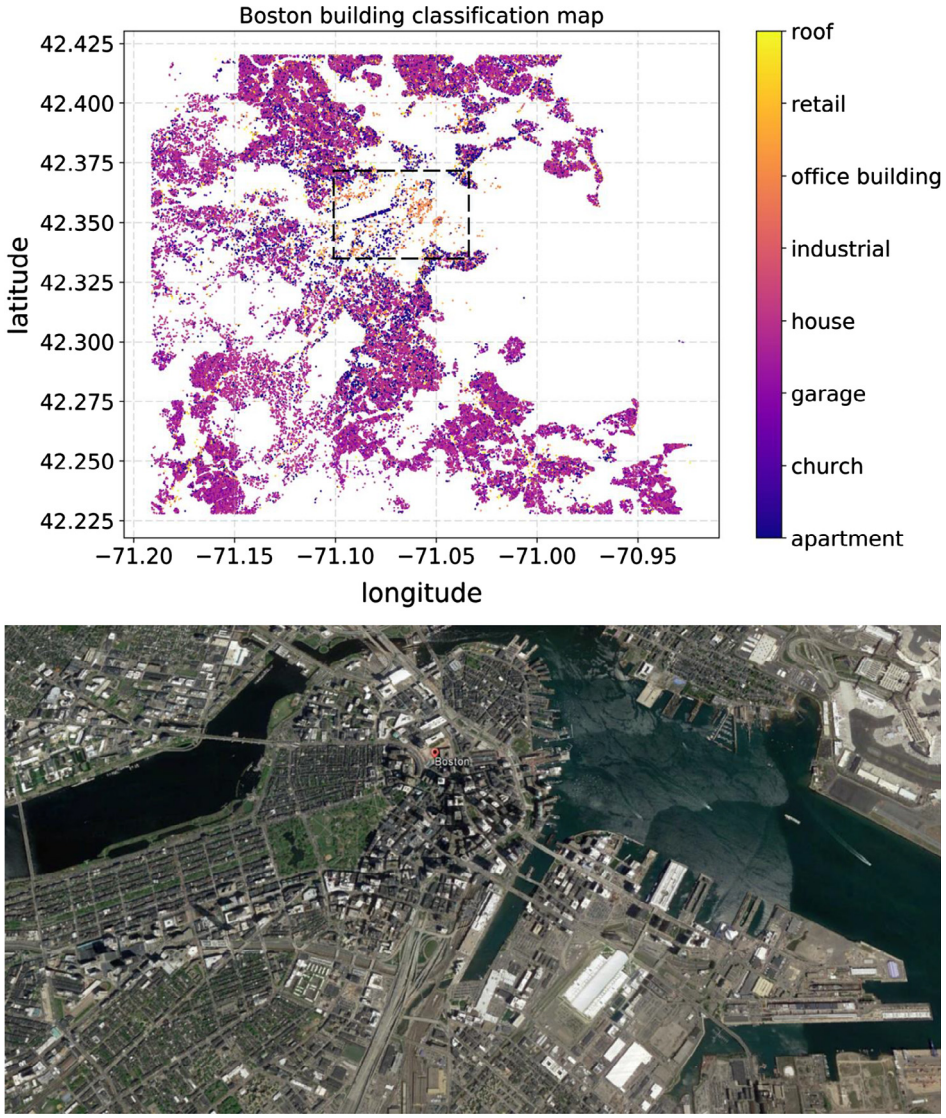

Fig. 21. The city-scale building classification map of Boston, where each classified building instance is displayed as a colored point with GPS coordinates. Since most office buildings and apartments locate in the cropped area, it can be observed that Boston is with one single central business district. It is not an industry city, as no large block of industrial districts is found. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

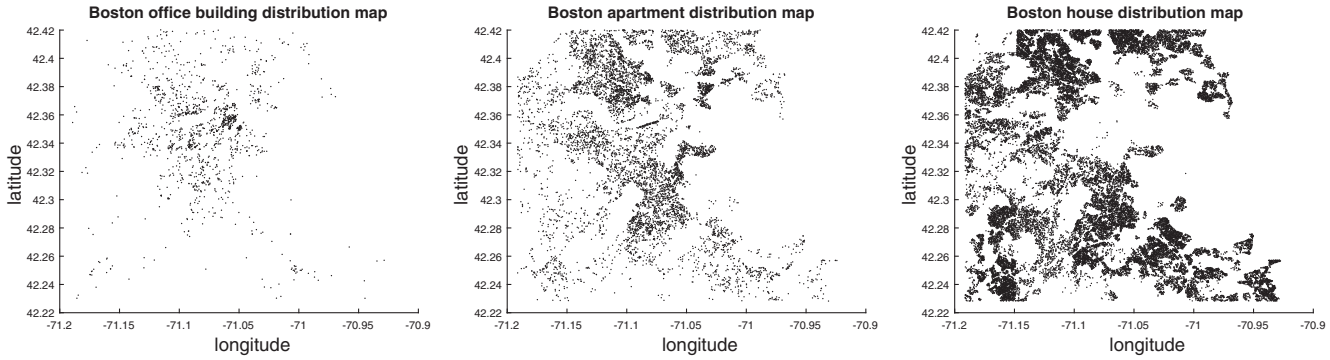

Fig. 22. Office building, apartment and house distribution maps of Boston. We can see that both the densities of office buildings and apartments decrease from the center to its outside, while it is contrary of the house density.

4.3.2. City-scale Maps of Calgary, Boston and Toronto

As shown in Figs. 19, 21 and 24, we provide the city-scale building classification maps of Calgary, Boston and Toronto based on classifying the retrieved 6124, 64,389 and 45,978 building street view images, respectively, where each classified building instance is displayed as a colored point with its GPS coordinate. Besides, Figs. 20,

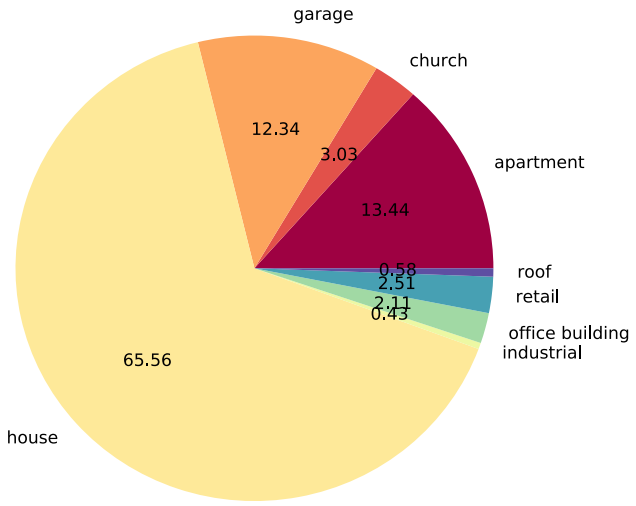

Fig. 23. Pie chart of the building class proportions of the predicted buildings of Boston.

23 and 26 demonstrate the associated numerical proportions of building classes based on the classification results. In order to quantitatively analyze the performance, 1000 buildings in each city are randomly selected and their associated building tags from OSM are retrieved according to their GPS locations. The classification performances of the three cities are demonstrated in Tables 3–5, respectively. Table 3 demonstrates that the overall accuracy of the classification result in Calgary is around 0.7, given the retrieved 1000 building tags from OSM. As illustrated in Table 4, by comparing with the building tags from OSM, the overall accuracy of the classification map in Boston can reach around 0.55. According to Table 5, more than 75% buildings in Toronto can be accurately classified.

As shown by the classification map of Calgary, it is obvious that there are three main industrial districts, and the downtown area is crowded by office buildings. Correspondingly, we also present the remote sensing images of one industrial and the downtown areas (black rectangles). Such classification map can infer that Calgary is an industry city with single central business district to which the three main industrial blocks are located.

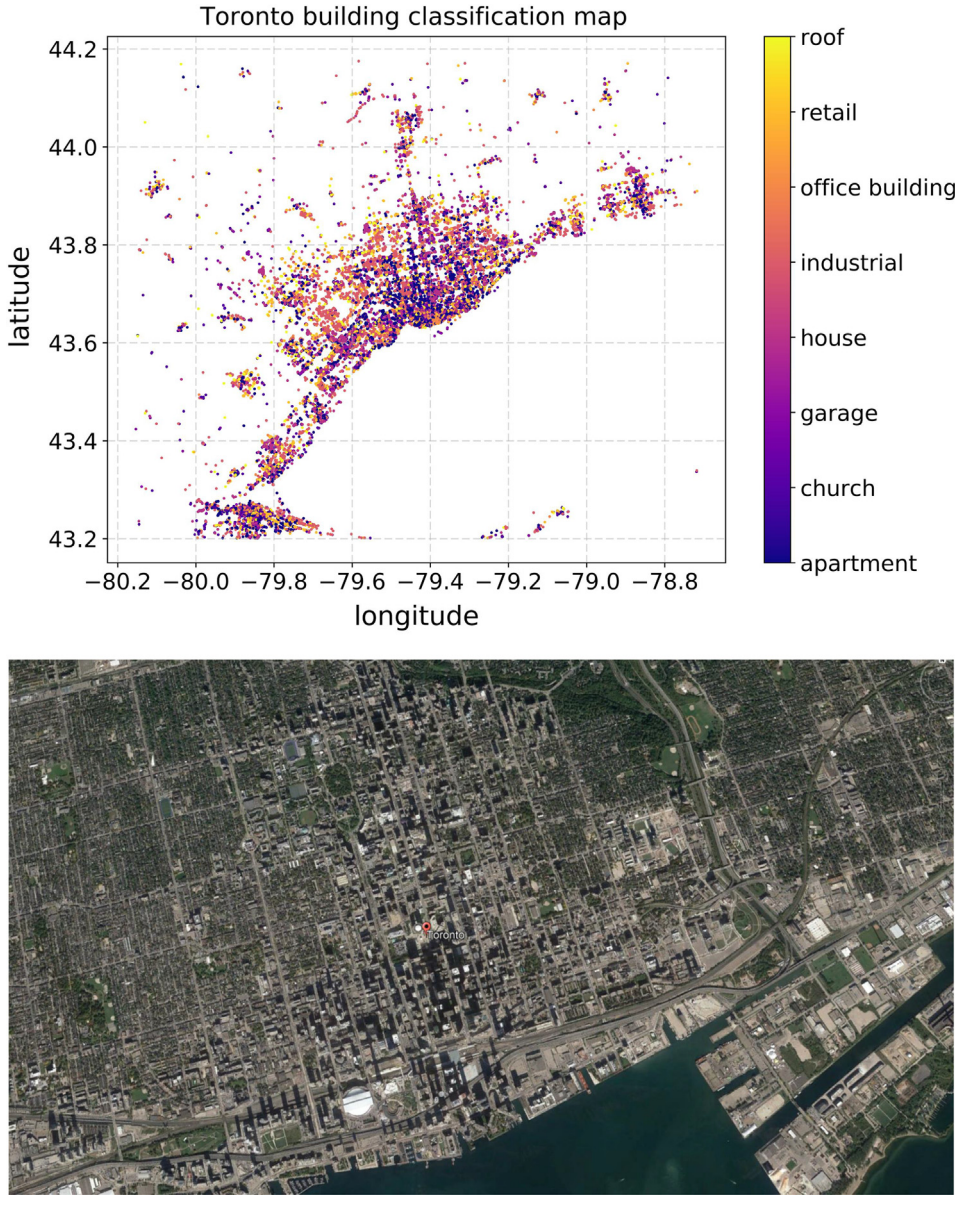

Fig. 24. The city-scale building classification map of Toronto and the associated remote sensing image of the central city. Most apartments and office buildings are located in the center of Toronto, and most industrial buildings are distributed in the regions around it.

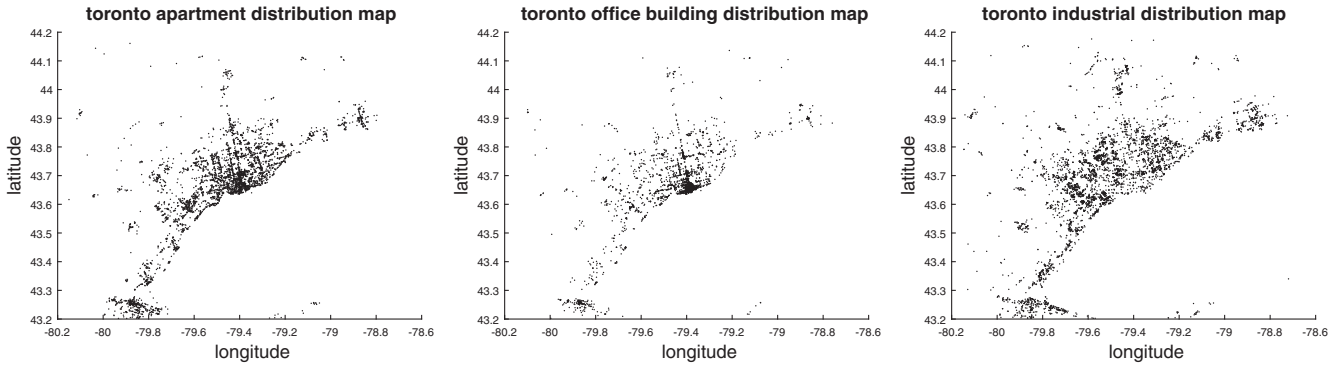

Fig. 25. Apartment, office and industrial building distribution maps of Toronto.

According to the map of Boston, houses obviously dominate the buildings of the city, and they are located around the city center. Besides, Boston is also with one single central business district (noted by the black dashed rectangle), since most office buildings and apartments locate in this area. The associated remote sensing image is demonstrated at the bottom of Fig. 21. As shown by the distribution maps of office buildings, houses and apartments plotted in Fig. 22, we can see that the densities of office buildings and apartments decrease from the center to its outside, while it is contrary of the house density. In addition, Boston is not an industry city, since no large block of industrial districts is observed in the classification map and the proportion of industrial buildings is very low.

From the maps of Fig. 24 and 25, most apartments and office buildings are located in the center of Toronto, and most industrial buildings are distributed in the regions around it. As shown in Fig. 26, the second largest proportion of building classes is apartment, which indicates that the population density is high in Toronto, especially in the city center. Besides, around 10% buildings are industrial, which indicates industry is one of the fields which contribute most to the economy of Toronto.

5. Discussion

In our training experiments, all the fully connected layers are initialized randomly. As for ResNet, there is only one fully connected layer (*softmax* layer) in the architecture and we utilize the network pre-trained on *ImageNet* dataset which contains totally 1000 classes. The parameters of the fully connected layer cannot be directly transferred to our task, since there are 8 classes in our dataset. While, for AlexNet and VGG16, besides the last fully connected layer (*softmax* layer), there are two more fully connected (fc) layers to be initialized. Taking VGG16 as an example, we took two experiments with the same hyperparameters for training the network on our benchmark dataset, while those two fully connected layers were initialized in two ways, i.e. initialized randomly and by the parameters pretrained with *ImageNet*. As shown in Fig. 27, VGG16 where the two fully connected layers were initialized by the pretrained parameters did accelerate the training of the network, since it can achieve higher classification accuracy than the one where the fc layers were initialized randomly during the first several epochs. However, both of them can converge to comparable classification accuracies at last.

According to the classification accuracies of eight building classes, churches are relatively easier to recognize than any other classes, since their structures are more unique, while some classes are not easily identified, e.g. retails and industrials. There are the following reasons that may influence the classification results. Firstly, since the ground truth labels come from the OSM users, manually labeling errors among some building classes exist in the benchmark dataset, especially for those with similar façade structures, e.g. some industrial and office buildings. As shown in Fig. 28(Left), the building displayed by the street view image tends to be an office building, while the building tag retrieved from OSM is industrial. Secondly, some street view images include multiple buildings of different classes, e.g. a house with a garage by its side.

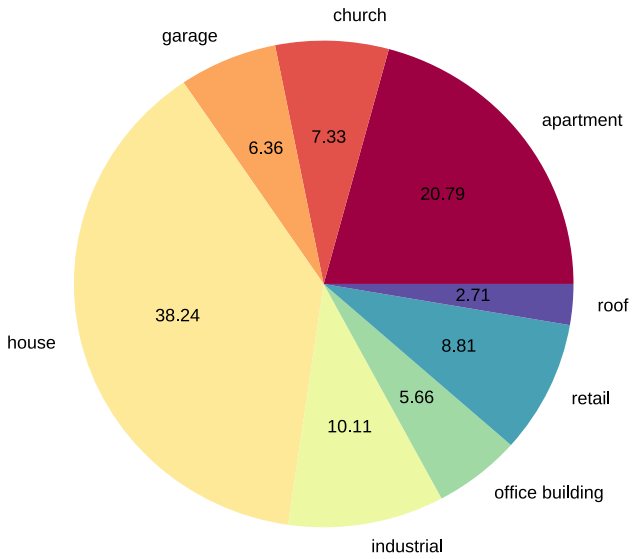

Fig. 26. Pie chart of the building class proportions of the predicted buildings of Toronto.

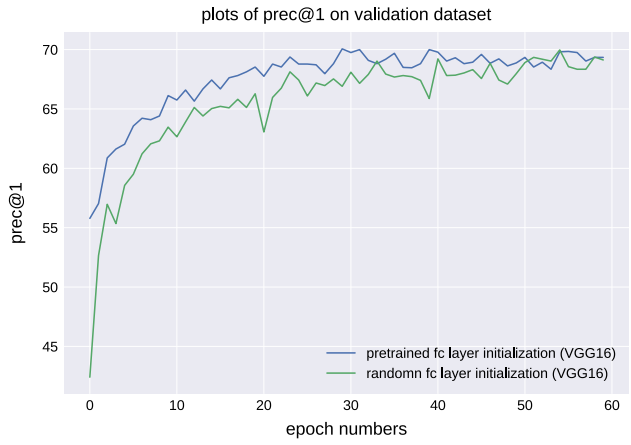

Fig. 27. Top-1 precision curves of VGG16 on the validation dataset with different initializations of fully connected layers.

Fig. 28. Some results show the reasons that may induce classification errors. (Left) The building displayed by the street view image tends to be an office building, while the ground truth building tag retrieved from OSM is industrial. (Middle) The building demonstrated in the image is a house, while it is misclassified as garage, since there are both garage and house structures demonstrated in the image. (Right) Although the building is correctly recognized, the confidence score is not so high, since the typical façade structure of church is not displayed in the retrieved image.

From Fig. 28(Middle), the building demonstrated in the image is a house, while it is misclassified as a garage. Lastly, side faces of buildings are displayed in some retrieved street view images, thus the corresponding façade features are not rich for the classification. As illustrated by Fig. 28(Right), although the building is correctly recognized, the confidence score is not so high, since the typical façade structure of church is not displayed in the retrieved image.

It is worth noting that as an alternative, a building rejection class can be added to replace the outlier removal procedure, depending on the quality of input data.

6. Conclusion and future work

In this paper, we presented a framework for building instance classification, which tended to provide more informative classification maps. With this approach, relatively high accuracies could be achieved for land-use classification of individual buildings. For this task, we built a street view benchmark dataset with eight building categories for training and testing. By investigating four different CNN architectures, we chose VGG16 to predict building instance classification maps on region and city scales. Such maps help us to get insight of urban areas, and have the potential for many innovative urban analysis, e.g. very high resolution urban population density mapping, urban social structure understanding, city economy structure analysis and general urban planning.

For the future work, to improve the classification performance, other information can be fused, e.g. text descriptions associated with social media images and text information displayed in images, e.g. brand names. Also, in order to obtain denser building classification maps, information from remote sensing images and GIS maps can be exploited for those buildings without street view images. In case that building footprints cannot be retrieved from GIS maps, a method of individual building detection in remote sensing images should be also developed.

Acknowledgment

We gratefully acknowledge the support of the European Research Council (ERC) under the European Unions Horizon 2020 research and innovation programme (grant agreement No [ERC-2016-StG-714087], Acronym: *So2Sat*), Helmholtz Association under the framework of the Young Investigators Group "SiPEO" (VH-NG-1018, www.sipeo.bgu.tum.de), the computing time granted by the John von Neumann Institute for Computing (NIC) and provided on the supercomputer JURECA at Jülich Supercomputing Centre (JSC), as well as NVIDIA Corporation with the donation of the Titan X Pascal GPU used in this research.

The authors would like to thank the reviewers for their valuable suggestions.

References

Albert, A., Kaur, J., Gonzalez, M., 2017. Using convolutional networks and satellite imagery to identify patterns in urban environments at a large scale. arXiv preprint <arXiv:1704.02965>.

J.R. Anderson, A land use and land cover classification system for use with remote sensor data, volume 964, US Government Printing Office, 1976.

Anguelov, D., Dulong, C., Filip, D., Frueh, C., Lafon, S., Lyon, R., Ogale, A., Vincent, L., Weaver, J., 2010. Google street view: capturing the world at street level. Computer 43, 32–38.

Chen, Y., Lin, Z., Zhao, X., Wang, G., Gu, Y., 2014. Deep learning-based classification of hyperspectral data. IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens. 7, 2094–2107.

Cheng, G., Han, J., Guo, L., Liu, Z., Bu, S., Ren, J., 2015. Effective and efficient midlevel visual elements-oriented land-use classification using VHR remote sensing images. IEEE Trans. Geosci. Remote Sens. 53, 4238–4249.

Cheng, G., Zhou, P., Han, J., 2016. Learning rotation-invariant convolutional neural networks for object detection in vhr optical remote sensing images. IEEE Trans. Geosci. Remote Sens. 54, 7405–7415.

Cheng, G., Han, J., Lu, X., 2017. Remote sensing image scene classification: benchmark and state of the art. Proc. IEEE.

Cheriyadat, A.M., 2014. Unsupervised feature learning for aerial scene classification. IEEE Trans. Geosci. Remote Sens. 52, 439–451.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE, pp. 248–255.

Di, K., Li, D., Li, D., 2000. Land use classification of remote sensing image with GIS data based on spatial data mining techniques. Int. Arch. Photogramm. Remote Sens. 33, 238–245.

Gong, P., Marceau, D.J., Howarth, P.J., 1992. A comparison of spatial feature extraction algorithms for land-use classification with spot HRV data. Remote Sens. Environ. 40, 137–151.

Han, J., Zhang, D., Cheng, G., Guo, L., Ren, J., 2015. Object detection in optical remote sensing images based on weakly supervised learning and high-level feature learning. IEEE Trans. Geosci. Remote Sens. 53, 3325–3337.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778.

Huang, X., Lu, Q., Zhang, L., 2014. A multi-index learning approach for classification of high-resolution remotely sensed images over urban areas. ISPRS J. Photogramm. Remote Sens. 90, 36–48.

Huang, X., Liu, H., Zhang, L., 2015. Spatiotemporal detection and analysis of urban villages in mega city regions of china using high-resolution remotely sensed imagery. IEEE Trans. Geosci. Remote Sens. 53, 3639–3657.

Huang, X., Wen, D., Li, J., Qin, R., 2017. Multi-level monitoring of subtle urban changes for the megacities of china using high-resolution multi-view satellite imagery. Remote Sens. Environ. 196, 56–75.

Hughes, L.H., Schmitt, M., Mou, L., Wang, Y., Zhu, X.X., 2018. Identifying corresponding patches in sar and optical images with a pseudo-siamese cnn. arXiv preprint <arXiv:1801.08467>.

Khorram, S., Brockhaus, J.A., Cheshire, H.M., 1987. Comparison of landsat MSS and tm data for urban land-use classification. IEEE Trans. Geosci. Remote Sens., 238–243

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft coco: Common objects in context. In: European Conference on Computer Vision. Springer, pp. 740–755.

Liu, Q., Hang, R., Song, H., Zhu, F., Plaza, J., Plaza, A., 2016. Adaptive deep pyramid matching for remote sensing scene classification. arXiv preprint <arXiv:1611.03589>.

Lowe, D.G., 1999. Object recognition from local scale-invariant features. The proceedings of the seventh IEEE international conference on Computer vision, 1999, vol. 2. IEEE, pp. 1150–1157.

Lu, D., Weng, Q., 2006. Use of impervious surface in urban land-use classification. Remote Sens. Environ. 102, 146–160.

Luus, F.P., Salmon, B.P., van den Bergh, F., Maharaj, B.T.J., 2015. Multiview deep learning for land-use classification. IEEE Geosci. Remote Sens. Lett. 12, 2448–2452.

Ma, X., Wang, H., Wang, J., 2016. Semisupervised classification for hyperspectral image based on multi-decision labeling and deep feature learning. ISPRS J. Photogramm. Remote Sens. 120, 99–107.

Maggiori, E., Tarabalka, Y., Charpiat, G., Alliez, P., 2017. Convolutional neural networks for large-scale remote-sensing image classification. IEEE Trans. Geosci. Remote Sens. 55, 645–657.

Marmanis, D., Datcu, M., Esch, T., Stilla, U., 2016. Deep learning earth observation classification using imagenet pretrained networks. IEEE Geosci. Remote Sens. Lett. 13, 105–109.

Mou, L., Ghamisi, P., Zhu, X.X., 2017. Deep recurrent neural networks for hyperspectral image classification. IEEE Trans. Geosci. Remote Sens. 55, 3639–3655.

Mou, L., Ghamisi, P., Zhu, X.X., 2018. Unsupervised spectral-spatial feature learning via deep residual conv-deconv network for hyperspectral image classification. IEEE Trans. Geosci. Remote Sens. 56, 391–406.

OpenStreetMap contributors, 2017. Planet dump retrieved from <https://planet.osm.org>, <https://www.openstreetmap.org>.

Pacifici, F., Chini, M., Emery, W.J., 2009. A neural network approach using multi-scale textural metrics from very high-resolution panchromatic imagery for urban land-use classification. Remote Sens. Environ. 113, 1276–1292.

Pal, M., Mather, P.M., 2003. An assessment of the effectiveness of decision tree methods for land cover classification. Remote Sens. Environ. 86, 554–565.

Paola, J.D., Schowengerdt, R.A., 1995. A detailed comparison of backpropagation neural network and maximum-likelihood classifiers for urban land use classification. IEEE Trans. Geosci. Remote Sens. 33, 981–996.

Penatti, O.A., Nogueira, K., dos Santos, J.A., 2015. Do deep features generalize from everyday objects to remote sensing and aerial scenes domains? In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 44–51.

Rigas, I., Economou, G., Fotopoulos, S., 2013. Low-level visual saliency with application on aerial imagery. IEEE Geosci. Remote Sens. Lett. 10, 1389–1393.

Rodriguez-Galiano, V.F., Ghimire, B., Rogan, J., Chica-Olmo, M., Rigol-Sanchez, J.P., 2012. An assessment of the effectiveness of a random forest classifier for land-cover classification. ISPRS J. Photogramm. Remote Sens. 67, 93–104.

Romero, A., Gatta, C., Camps-Valls, G., 2016. Unsupervised deep feature extraction for remote sensing image classification. IEEE Trans. Geosci. Remote Sens. 54, 1349–1362.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al., 2015. Imagenet large scale visual recognition challenge. Int. J. Comput. Vision 115, 211–252.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint <arXiv:1409.1556>.

Stefanov, W.L., Ramsey, M.S., Christensen, P.R., 2001. Monitoring urban land cover change: an expert system approach to land cover classification of semiarid to arid urban centers. Remote Sens. Environ. 77, 173–185.

Sun, X., Nasrabadi, N.M., Tran, T.D., 2015. Task-driven dictionary learning for hyperspectral image classification with structured sparsity constraints. IEEE Trans. Geosci. Remote Sens. 53, 4457–4471.

Taubenböck, H., Klotz, M., Wurm, M., Schmieder, J., Wagner, B., Wooster, M., Esch, T., Dech, S., 2013. Delineation of central business districts in mega city regions using remotely sensed data. Remote Sens. Environ. 136, 386–401.

Tuia, D., Flamary, R., Courty, N., 2015. Multiclass feature learning for hyperspectral image classification: Sparse and hierarchical solutions. ISPRS J. Photogramm. Remote Sens. 105, 272–285.

Tuia, D., Flamary, R., Barlaud, M., 2016. Nonconvex regularization in remote sensing. IEEE Trans. Geosci. Remote Sens. 54, 6470–6480.

Wang, Z., Nasrabadi, N.M., Huang, T.S., 2014. Spatial–spectral classification of hyperspectral images using discriminative dictionary designed by learning vector quantization. IEEE Trans. Geosci. Remote Sens. 52, 4808–4822.

Xu, X., Li, J., Huang, X., Dalla Mura, M., Plaza, A., 2016. Multiple morphological component analysis based decomposition for remote sensing image classification. IEEE Trans. Geosci. Remote Sens. 54, 3083–3102.

Yang, Y., Newsam, S., 2010. Bag-of-visual-words and spatial extensions for land-use classification. In: Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems. ACM, pp. 270–279.

Yang, S., Jin, H., Wang, M., Ren, Y., Jiao, L., 2014. Data-driven compressive sampling and learning sparse coding for hyperspectral image classification. IEEE Geosci. Remote Sens. Lett. 11, 479–483.

Yao, X., Han, J., Cheng, G., Qian, X., Guo, L., 2016. Semantic annotation of high-resolution satellite images via weakly supervised learning. IEEE Trans. Geosci. Remote Sens. 54, 3660–3671.

Yuan, F., Sawaya, K.E., Loeffelholz, B.C., Bauer, M.E., 2005. Land cover classification and change analysis of the twin cities (minnesota) metropolitan area by multitemporal landsat remote sensing. Remote Sens. Environ. 98, 317–328.

Zhang, F., Du, B., Zhang, L., 2015. Saliency-guided unsupervised feature learning for scene classification. IEEE Trans. Geosci. Remote Sens. 53, 2175–2184.

Zhang, L., Zhang, L., Du, B., 2016. Deep learning for remote sensing data: a technical tutorial on the state of the art. IEEE Geosci. Remote Sens. Mag. 4, 22–40.

Zhang, C., Pan, X., Li, H., Gardiner, A., Sargent, I., Hare, J., Atkinson, P.M., 2018. A hybrid mlp-cnn classifier for very fine resolution remotely sensed image classification. ISPRS J. Photogramm. Remote Sens. 140, 133–144.

Zhao, W., Du, S., 2016. Learning multiscale and deep representations for classifying remotely sensed imagery. ISPRS J. Photogramm. Remote Sens. 113, 155–165.

Zhou, B., Khosla, A. Lapedriza, A., Torralba, A., Oliva, A., 2016. Places: an image database for deep scene understanding, arXiv preprint <arXiv:1610.02055>.

Zhu, Q., Zhong, Y., Zhao, B., Xia, G.-S., Zhang, L., 2016. Bag-of-visual-words scene classifier with local and global features for high spatial resolution remote sensing imagery. IEEE Geosci. Remote Sens. Lett. 13, 747–751.

Zhu, X.X., Tuia, D., Mou, L., Xia, G.S., Zhang, L., Xu, F., Fraundorfer, F., 2017. Deep learning in remote sensing: a comprehensive review and list of resources. IEEE Geosci. Remote Sens. Mag. 5, 8–36.

Zou, Q., Ni, L., Zhang, T., Wang, Q., 2015. Deep learning based feature selection for remote sensing scene classification. IEEE Geosci. Remote Sens. Lett. 12, 2321–2325.